LREC-COLING 2024

TRAC-2024: The Fourth Workshop on Threat, Aggression & Cyberbullying @LREC-COLING-2024

Workshop Proceedings

Editors
Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, Bharathi Raja Chakravarthi, Bornini Lahiri, Siddharth Singh and Shyam Ratan

20 May, 2024
Torino, Italia

Proceedings of the TRAC-2024: The Fourth Workshop on Threat, Aggression & Cyberbullying @LREC-COLING-2024

ISBN 978-2-493814-47-0
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Introduction

As the number of users and their web-based interaction has increased, incidents of verbal threat, aggression and related behavior like trolling, cyberbullying, and hate speech have also increased manifold globally. The reach and extent of the Internet have given such incidents unprecedented power and influence to affect the lives of billions of people. Such incidents of online abuse have not only resulted in mental health and psychological issues for users, but they have manifested in other ways, spanning from deactivating social media accounts to instances of self-harm and suicide and offline violence as well.
To mitigate these issues, researchers have begun to explore the use of computational methods for identifying such toxic interactions online. In particular, Natural Language Processing (NLP) and ML-based methods have shown great promise in dealing with such abusive behaviour through early detection of inflammatory content. In fact, we have observed an explosion of NLP-based research on offensive content in the last few years. The creation of new venues such as the WOAH and the TRAC workshop series has accompanied this growth. Community-based competitions, like tasks 5/6 at SemEval-2019, task 12 at SemEval-2020, task 5/7 at SemEval-2021, task 7 at SemEval-2023 have also proven extremely popular. In fact, because of the huge community interest, multiple workshops are being held on the topic in a single year. For example, in 2018 ACL hosted both the Abusive Language Online workshop (EMNLP) as well as TRAC-1 (COLING). Both venues achieved healthy participation with 21 and 24 papers, respectively. Interest in the topic has continued to grow since then.

We understand that a synergy and mutual cooperation needs to be established between the linguistic analysis of impolite, threatening, aggressive and hateful language (from pragmatic, sociolinguistic, discourse analysis and other perspectives) and NLP and ML (including deep learning) - based approaches to identification of such languages. As such we actively focus on bringing the two communities together to develop a better understanding of these issues. The workshop provides a forum for everyone working in the area to discuss their research and for further collaboration. We proposed a new edition of the workshop to support the community and further research in this area.

As in the earlier editions, TRAC focuses on the applications of NLP, ML and pragmatic studies on aggression and impoliteness to tackle these issues. As such the workshop also includes a shared task on **"HarmPot-ID: Offline Harm Potential Identification"**. It has introduced the novel task of predicting the offline harm potential of social media posts - broadly the task is to predict whether a specific post is likely to initiate, incite or further exaggerate an offline harm event (viz. riots, mob lynching, murder, rape, etc). It consisted of two sub-tasks.

- **Sub-task 1a: What is the offline harm potential of a document?:** It was a four-class classification task where the participants were required to predict the level of offline harm potential -
  - 0 (it will never lead to offline harm, in any context),
  - 1 (it could lead to incite an offline harm event given specific conditions or context),
  - 2 (it is most likely to incite in most contexts or probably initiate an offline harm event in specific contexts)
  - 3 (it is certainly going to incite or initiate an offline harm event in any context).

- **Sub-task 1b: Who is/are the most likely target(s) of the offline harm?:** If an offline harm event is triggered, who are going to be the most affected groups of people? In this task, only the broad category of the target(s) identities are to be predicted. It was a five-class classification task - Gender, Religion, Descent, Caste and Political Ideology

Both the workshop and the shared task received a very encouraging response from the community. The proceedings include 9 oral and 8 posters (including 3 system description papers). We would like to thank all the authors for their submissions and members of the Program Committee for their invaluable efforts in reviewing and providing feedback to all the papers. We would also like to thank all the members of the Organising Committee who have helped immensely in various aspects of the organisation of the workshop and the shared task.

Workshop Chairs

Workshop Chairs

Ritesh Kumar, Council for Strategic and Defense Research, India and UnReaL-TecE LLP, India
Atul Kr. Ojha, University of Galway, Ireland & Panlingua Language Processing LLP, India
Shervin Malmasi, Amazon USA
Bharathi Raja Chakravarthi, University of Galway, Ireland
Bornini Lahiri, Indian Institute of Technology, Kharagpur, India
Siddharth Singh, UnReaL-TecE LLP, India
Shyam Ratan, UnReaL-TecE LLP, India

Programme Committee

Anagha HC, National Institute of Technology Karnataka
Arup Baruah, Indian Institute of Information Technology, Guwahati
Atul Kr. Ojha, University of Galway, Ireland & Panlingua, India
Bornini Lahiri, Indian Institute of Technology-Kharagpur, India
Bruno Emanuel Martins, IST and INESC-ID
Chuan-Jie Lin, National Taiwan Ocean University, Taiwan
Denis Gordeev, The Russian Presidential Academy of National Economy and Public Administration under the President of the Russian Federation
Ilia Markov, Vrije Universiteit Amsterdam, CLTL
Jack Depp, Nanjing University of Science and Technology
Kirti Kumari, National Institute of Technology Patna
Lee Gillam, University of Surrey
Manuel Montes-y-Gómez, INAOE, Mexico
Marcos Zampieri, George Mason University
Min-Yuh Day, Tamkang University
Nemanja Djuric, Aurora Innovation
Parth Patwa, University of California Los Angeles
Ritesh Kumar, Council for Strategic and Defense Research, India and UnReaL-TecE LLP, India
Ruifeng Xu, Harbin Institute of Technology, China
Saja Tawalbeh, University of Antwerp
Sarang Gupta, Columbia University
Shubhanshu Mishra, Twitter Inc.
Shyam Ratan, UnReaL-TecE LLP, India
Siddharth Singh, UnReaL-TecE LLP, India
Valerio Basile, University of Turin
Yeshan Wang, Vrije Universiteit Amsterdam
Zeerak Talat, Independent Researcher

Table of Contents

Conference Program

Monday, May 20, 2024

09:00–09:10 **Inaugural Session**
Chair: Workshop Chairs

09:00–09:10 *Welcome*
Workshop Chairs

09:10–10:00 **Keynote Talk**
Chair: Bharathi Raja Chakravarthi

10:00–10:30 **Oral Session-I**
Chair: Bharathi Raja Chakravarthi

10:00–10:30 *The Constant in HATE: Toxicity in Reddit across Topics and Languages*
Wondimagegnhue Tsegaye Tufa, Ilia Markov and Piek T.J.M. Vossen

10:30–11:00 ***Cofee Break and Poster Session***

10:30–11:00 *A Federated Learning Approach to Privacy Preserving Offensive Language Identification*
Marcos Zampieri, Damith Premasiri and Tharindu Ranasinghe

10:30–11:00 *CLTL@HarmPot-ID: Leveraging Transformer Models for Detecting Offline Harm Potential and Its Targets in Low-Resource Languages*
Yeshan Wang and Ilia Markov

10:30–11:00 *NJUST-KMG at TRAC-2024 Tasks 1 and 2: Offline Harm Potential Identification*
Jingyuan Wang, Jack Depp and Yang Yang

10:30–11:00 *ScalarLab@TRAC2024: Exploring Machine Learning Techniques for Identifying Potential Offline Harm in Multilingual Commentaries*
Anagha H C, Saatvik M. Krishna, Soumya Sangam Jha, Vartika T. Rao and Anand Kumar M

Monday, May 20, 2024 (continued)

11:00–13:00 Oral Session-II

11:00–11:30 *LLM-Based Synthetic Datasets: Applications and Limitations in Toxicity Detection*
Udo Kruschwitz and Maximilian Schmidhuber

11:30–12:00 *Using Sarcasm to Improve Cyberbullying Detection*
Xiaoyu Guo and Susan Gauch

12:00–12:30 *Analyzing Offensive Language and Hate Speech in Political Discourse: A Case Study of German Politicians*
Maximilian Weissenbacher and Udo Kruschwitz

12:30–13:00 *Ice and Fire: Dataset on Sentiment, Emotions, Toxicity, Sarcasm, Hate speech, Sympathy and More in Icelandic Blog Comments*
Steinunn Rut Friðriksdóttir, Annika Simonsen, Atli Snær Ásmundsson, Guðrún Lilja Friðjónsdóttir, Anton Karl Ingason, Vésteinn Snæbjarnarson and Hafsteinn Einarsson

13:00–14:00 *Lunch Break*

14:00–15:00 Oral Session-III

14:00–14:30 *Detecting Hate Speech in Amharic Using Multimodal Analysis of Social Media Memes*
Melese Ayichlie Jigar, Abinew Ali Ayele, Seid Muhie Yimam and Chris Biemann

14:30–15:00 *Content Moderation in Online Platforms: A Study of Annotation Methods for Inappropriate Language*
Baran Barbarestani, Isa Maks and Piek T.J.M. Vossen

Monday, May 20, 2024 (continued)

15:00–16:00 **Panel Discussion**
Chair: TBD

16:00–16:30 ***Cofee Break***

16:00–16:30 *FrenchToxicityPrompts: a Large Benchmark for Evaluating and Mitigating Toxicity in French Texts*
Caroline Brun and Vassilina Nikoulina

16:00–16:30 *Studying Reactions to Stereotypes in Teenagers: an Annotated Italian Dataset*
Elisa Chierchiello, Tom Bourgeade, Giacomo Ricci, Cristina Bosco and Francesca D'Errico

16:00–16:30 *Offensiveness, Hate, Emotion and GPT: Benchmarking GPT3.5 and GPT4 as Classifiers on Twitter-specific Datasets*
Nikolaj Bauer, Moritz Preisig and Martin Volk

16:00–16:30 *DoDo Learning: Domain-Demographic Transfer in Language Models for Detecting Abuse Targeted at Public Figures*
Angus Redlarski Williams, Hannah Rose Kirk, Liam Burke-Moore, Yi-Ling Chung, Ivan Debono, Pica Johansson, Francesca Stevens, Jonathan Bright and Scott Hale

16:30–17:30 **Oral Session-IV**

16:30–17:00 *Empowering Users and Mitigating Harm: Leveraging Nudging Principles to Enhance Social Media Safety*
Gregor Donabauer, Emily Theophilou, Francesco Lomonaco, Sathya Bursic, Davide Taibi, Davinia Hernández-Leo, Udo Kruschwitz and Dimitri Ognibene

17:00–17:30 *Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse*
Abinew Ali Ayele, Esubalew Alemneh Jalew, Adem Chanie Ali, Seid Muhie Yimam and Chris Biemann

Monday, May 20, 2024 (continued)

17:30–17:40 Closing

17:30–17:40 *Vote of Thanks*
Workshop Chairs